ASHA American Speech-Language-Hearing Association

AJA

Research Note

Feasibility of Tablet-Based Remote Data Collection Method for Measuring Hearing Aid Preference

Varsha Rallapalli[a] and **Pamela Souza**[a,b]

[a] Roxelyn and Richard Pepper Department of Communication Sciences & Disorders, Northwestern University, Evanston, IL [b] Knowles Hearing Center, Northwestern University, Evanston, IL

ARTICLE INFO

Article History:
Received December 16, 2021
Revision received April 4, 2022
Accepted May 2, 2022

Editor-in-Chief: Ryan W. McCreery
Editor: Erin M. Picou

https://doi.org/10.1044/2022_AJA-21-00273

ABSTRACT

Purpose: The purpose of this study was to determine the feasibility of a tablet-based remote data collection method for measuring preference for hearing aid signal processing features.
Method: Participants were nine individuals with bilateral mild to moderately severe sensorineural hearing loss. Stimuli were spatialized low-context sentences mixed with six-talker babble at two realistic signal-to-noise ratios (3 and 8 dB) and processed through a hearing aid simulator. Preference for full factorial combinations of three common hearing aid processing features (two levels each) was elicited using a paired-comparison task. Participants completed two versions of the experiment: The lab version was completed in a sound-treated booth using a custom MATLAB application on a desktop computer; the remote version was completed in a quiet room in the participant's home, using a custom MATLAB executable application on a tablet. Both versions used the same calibrated headphones. Strict infection control protocols were followed.
Results: McNemar's test showed no association between preference and data collection method for the majority of the conditions. Percentage agreement and kappa scores were moderate/fair across most conditions. The results indicated that the remote versus lab versions did not have a systematic effect on preference. However, the relatively low agreement and kappa scores suggested within-subject variability in the outcome (preference).
Conclusion: The tablet-based version of remote experimentation was comparable to the lab-based version for eliciting preference for hearing aid signal processing features.

Researchers have long had interests in improved systems for remote data collection, both to support broader recruitment and to remove barriers to in-lab participation for some groups (e.g., Schoeffler et al., 2018; Strori et al., 2020; Walker & Campbell-Kibler, 2015; Yu & Lee, 2014). The need for remote systems was abruptly increased by the COVID-19 pandemic and several audiology research labs had to consider alternative means to conduct experiments (e.g., Kopun et al., 2021; McPherson & McDermott, 2020; Waz & Chubb, 2020). For example, the Acoustical Society of America Psychological and Physiological Acoustics Task Force on Remote Testing (2020) identified several browser-based (e.g., Gorilla, Amazon Mechanical Turk), tablet-based (e.g., in-house custom applications), and desktop platforms (e.g., Research Electronic Data Capture or REDCap; Harris et al., 2009, 2019) that may be used to design and conduct such experiments. Even before the pandemic, browser-based platforms were successfully used to collect remote data for psychoacoustic experiments involving individuals with normal hearing (e.g., Eskenazi et al., 2013; Strori et al., 2020). While browser-based platforms are advantageous for crowdsourcing and recruiting a diverse population, existing methods pose several challenges for conducting perceptual experiments with individuals with hearing loss. These challenges include the inability to remotely calibrate participant's personal equipment, poor signal fidelity and limited headroom with personal

Correspondence to Varsha Rallapalli: varsha.rallapalli@northwestern.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

headphones, inability to verify the listener's degree of hearing loss, and uncertainty about background noise levels. Moreover, for experiments applying hearing aid signal processing that need to be customized to the listener, variations in headroom and difficulty in reliably calibrating personal headphones are particularly disadvantageous. This is because limiting gain due to a low headroom can result in poor audibility of the signals, whereas too much gain may be uncomfortable for the listener or result in a distorted signal due to peak clipping (Fortune & Preves, 1992; Souza, 2002). In order to isolate the effects of specific hearing aid signal processing on listeners with hearing loss, we also need to avoid confounds due to differences in audibility across individuals and ensure that everyone receives frequency-specific gain according to their hearing levels. Thus, there is a need to evaluate a method of remote deployment that can overcome the abovementioned constraints for experiments that incorporate hearing aid signal processing. A tablet-based platform with calibrated headphones that can be safely delivered to a participant may be a viable option for this purpose. There is precedent for successful application of calibrated tablet-based platforms in research involving surveys, hearing tests, speech recognition tests, and background noise measurements (e.g., Shapiro et al., 2020). However, systems for measuring outcomes with hearing aid signal processing have not yet been evaluated.

This study was conducted to determine the feasibility of a tablet-based remote data collection method for measuring preference for hearing aid signal processing. This is part of a study with a larger group of listeners designed to elicit preference for combinations of hearing aid signal processing. The rationale for that study is briefly summarized here.

Evidence suggests that hearing aid signal processing directed at improving access to speech sounds may do so at the cost of degrading other important cues in the speech signal. For instance, faster wide dynamic range compression (WDRC) systems improve the audibility of the signal but also introduce temporal envelope distortions (Alexander & Masterson, 2015; Jenstad & Souza, 2005; Moore, 2008). Similarly, stronger frequency compression (FC) parameters trade audibility of high-frequency consonant information for the alteration of lower frequency vowels and formant transitions (Alexander, 2016b; Souza et al., 2013). Stronger digital noise reduction (DNR) algorithms successfully reduce noisy portions of the signal, at the cost of removing coincident speech information (Arehart et al., 2015; Bentler & Chiou, 2006; Brons et al., 2015). In turn, the choice of specific hearing aid signal processing settings can affect perceptual outcomes for hearing aid users (Alexander, 2016b; Kim & Loizou, 2011; Souza, Arehart, & Neher, 2015; Souza et al., 2019). Inappropriate selection of hearing aid settings can result in poor speech and quality perception and consequently lead to dissatisfaction and low rates of hearing aid adoption (Kochkin, 2007; McCormack & Fortnum, 2013). While research on perceptual outcomes with hearing aid signal processing settings has predominantly focused on speech intelligibility and quality, less attention has been given to the hearing aid users' preferences. Moreover, there is a lack of understanding about how multiple signal processing settings influence the user's preference when they are activated together and to different extents in hearing aids (as is the case in wearable hearing aids). Understanding a user's preference may improve customization during the hearing aid fitting and potentially improve user satisfaction (Amlani & Schafer, 2009; Kuk, 2002). Relevant to this study, measurement of preference is particularly amendable to remote delivery because it does not require the listener to repeat or type out responses, as they might for an open-set speech intelligibility test. Moreover, recent studies have shown that ambient noise in home environments is low (Kopun et al., 2021; Ramos et al., 2022) and therefore not likely to interfere with a listening task for individuals with hearing loss using supra-aural headphones.

This study compares preference (using a paired-comparison task) in a group of individuals with hearing loss who completed two versions of the experiment, each in a different location: One version was completed in the laboratory prior to the pandemic, and another version was completed remotely during the pandemic. Because this study was focused on eliciting preference for hearing aid signal processing in individuals with hearing loss, we chose a tablet-based platform with a custom MATLAB executable application for the remote version. This allowed us to replicate several aspects of the laboratory version of the experiment and minimize the methodological differences due to the test environment and equipment.

Method

Participants

Ten individuals (eight men and two women) with bilateral mild-to-moderately severe sensorineural hearing loss participated in the study. These 10 individuals were selected because they participated in the laboratory version of the study as it was originally designed. Participants were in the age range of 54–86 years (M = 70.4 years). Pure-tone audiograms were previously obtained in a clinical environment (double-walled sound-treated booth). Air-conduction thresholds were obtained at octave and mid-octave frequencies between 250 and 8000 Hz. Bone-conduction thresholds were obtained at octave frequencies between 500 and 4000 Hz. All participants had symmetric

hearing (asymmetry was defined as a difference of at least 15 dB HL at two or more frequencies or a difference of at least 20 dB HL at one frequency between 250 and 3000 Hz) and air–bone gaps of less than 15 dB HL at all octave frequencies. At the time of remote data collection, seven participants had an audiogram that was completed more than 7 months ago. For these participants, air-conduction thresholds were repeated using a remotely deployed validated automated audiometer (Grason-Stadler Automated Method for Testing Auditory Sensitivity FLEX; Margolis et al., 2016; Mosley et al., 2019) and confirmed to be within ±10 dB across test frequencies. All participants were native English speakers, had normal cognitive functioning based on the Montreal Cognitive Assessment (Nasreddine et al., 2005; > 22; Shen et al., 2016), and reported good health. Participants completed an informed consent process approved by Northwestern University's Institutional Review Board.

Participants were instructed to remove their personal hearing aids (if any) during the experiment because amplified signals were delivered via headphones. However, during the in-lab experiment, one participant forgot to remove their personal hearing aids. Data from this participant were excluded. Thus, nine participants were included in the final data analysis. Figure 1 shows the air-conduction thresholds obtained in the clinical environment for these nine participants.

Stimuli

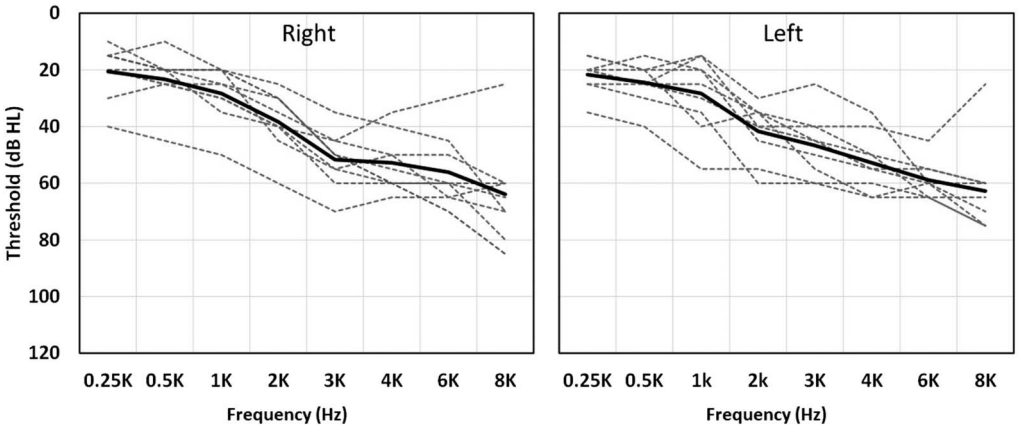
Stimuli were sentences mixed with multitalker babble. Target sentences were randomly sourced from the Institute of Electrical and Electronics Engineers (IEEE) database (Rothauser et al., 1969) and were spoken by two male and two female local talkers (Panfili et al., 2017). Multitalker babble consisted of six talkers from the same IEEE database but spoken by a different set of three male and three female talkers. To create the babble for a single trial, three randomly selected sentences (different from the target sentence) by a given talker were concatenated without any gaps. The concatenated string of sentences was then sliced to match the duration of the target sentence along with 1-s lead and lag times. To maintain a realistic scenario, no relative level adjustments were made among the six babble talkers. The target sentences and babble were mixed at two signal-to-noise ratios (SNRs): +3 and +8 dB. These SNRs were selected to represent the range of real-world SNRs (Smeds et al., 2015; Wu et al., 2018). Onset and offset ramps of 250 ms each were applied to the overall babble prior to mixing with the target sentence. The level of the sentence was fixed at 65 dB SPL, and the overall level of the six-talker babble was adjusted to match the desired SNR.

Prior to amplification processing, a room simulator (Zahorik, 2009) was used to spatialize the target speech and the six babble talkers under headphones. Binaural impulse responses (BRIRs) were generated for a small room (dimensions: 5.67 × 4.26 × 2.58 m) with absorption coefficients representative of an anechoic chamber ($\alpha = 1.0$ across all frequencies) and the listener seated in the center. The target sound source was located directly in front of the listener at 0° azimuth, and the six individual babble talkers were equally spaced around the listener between 90° and 270° azimuth and were randomly assigned to the locations across trials. All seven sound sources were placed at an equal distance of 1 m from the listener. The BRIRs were convolved with each sound source and then mixed at a given SNR to generate the speech in noise signals.

Hearing Aid Processing

The spatialized stimuli were processed through a hearing aid simulator in MATLAB (Arehart et al., 2015;

Figure 1. Air-conduction thresholds of nine participants included in the final analyses. Solid lines are the average for each ear across participants.



Kates et al., 2019; Souza, Arehart, Shen, et al., 2015). The hearing aid processing was a linear-phase 12-channel filter bank with center frequencies at 125, 250, 375, 500, 750, 1000, 1500, 2000, 3000, 4000, 5000, and 6000 Hz. Input to the hearing aid simulator was filtered through a microphone response for a behind-the-ear hearing aid shell (Arehart et al., 2022). Within a given channel in the hearing aid simulator, the first stage was nonlinear FC, followed by DNR, and finally WDRC. Parameters for each stage of signal processing are representative of the range available in clinical hearing aids (Dillon, 2012; Rallapalli et al., 2018) and are described below. The simulator had an overall delay of 10 ms to match the average delays found in commercial hearing aids (Alexander, 2016a). The output of the hearing aid was filtered through a typical receiver response as described in Arehart et al. (2021). No venting was applied.

For the WDRC stage of processing, the compression threshold was set to 45 dB SPL and inputs below this level received linear amplification. Inputs beyond an upper compression threshold of 100 dB SPL were subjected to compression limiting. Individualized compression ratios (CRs) were set across frequencies based on the NAL-NL2 procedure (Keidser et al., 2012). Two WDRC speed conditions were created: fast-acting (Fast) and slow-acting (Slow) with release times of 50 ms and 2,000 ms, respectively. The attack time was set to 5 ms for both conditions.

Nonlinear FC was simulated using sinusoidal modeling as described in Souza, Arehart, Shen, et al. (2015). Two FC conditions were created: OFF and ON. When FC was ON, the start frequency (SF) and CR were determined by typical settings used in commercial hearing aids for a given degree of hearing loss and audiometric configuration (Arehart et al., 2021). Seven participants had an SF of 4.7 kHz and a CR of 2.7. The remaining two participants with comparatively greater degrees of hearing loss had an SF of 4.6 kHz and a CR of 2.8.

DNR was implemented through Wiener filtering (Kates, 2017). The Cohen and Berdugo (2002) algorithm estimated the noise using a peak detector with 10-ms attack time and 100-ms release time. The algorithm classified the signal as speech versus noise based on the overall probability of the signal envelope exceeding a certain threshold. The noise power in each frequency band was computed for the duration of the stimulus, and the overall average was used as the noise power estimate for all the speech segments in that band. Once the signal was classified as noise and the noise power was estimated, Wiener filtering was applied to attenuate (or suppress) the noisy speech segments within a band. Two DNR conditions were used: low DNR and high DNR with a maximum attenuation of 3 dB and 12 dB, respectively.

Individualized gain across frequencies was provided for each participant based on the NAL-NL2 prescriptive method (Keidser et al., 2012). As all participants had symmetric hearing, the individualized frequency response for both ears was based on their better ear thresholds between 250 and 4000 Hz. If the thresholds were perfectly matched in both ears, the frequency response was based on the right ear. Additional details about the hearing aid simulator settings are provided in Appendix.

Preference

Listener preference was elicited using an adaptation of the choice-based conjoint analysis approach which typically involves five stages: (a) identifying attributes, (b) assigning levels, (c) formulating scenarios, (d) establishing preferences, and (e) data analysis (Bridges et al., 2012). In the choice-based conjoint analysis method, attributes are determined based on a rigorous qualitative analysis. However, this study modified this approach to use preselected hearing aid signal processing as attributes, including levels of WDRC, FC, and DNR, as described above that are commonly engaged in commercial hearing aids (Dillon, 2012; Rallapalli et al., 2018; Ricketts et al., 2019). Each attribute was assigned two levels: mild and strong. Specifically, the mild levels corresponded to Slow WDRC, Low DNR, or FC OFF and the strong levels corresponded to Fast WDRC, High DNR, or FC ON.

Based on the experimental design of choice-based conjoint analysis, preference was elicited using a paired-comparison task for auditory stimuli. A set of attributes, that is, one level of each signal processing technique represented a "Hearing Aid." A full-factorial design of all the attributes resulted in eight possible "Hearing Aids" (see Table 1). SNR was presented as a blocking variable. This resulted in 112 possible pairwise comparisons (64 "Hearing Aid" pairs × 2 SNRs). Out of the 64 "Hearing Aid" pairs, eight were identical pairs and were excluded, resulting in 56 pairs for analysis in this study.

Table 1. Combination of signal processing attributes and levels within each that constitute the eight "Hearing Aids."

"Hearing Aid"	Signal processing attributes & levels
S1	Slow WDRC, low DNR, FC OFF
S2	Slow WDRC, low DNR, FC ON
S3	Slow WDRC, high DNR, FC OFF
S4	Slow WDRC, high DNR, FC ON
S5	Fast WDRC, low DNR, FC OFF
S6	Fast WDRC, low DNR, FC ON
S7	Fast WDRC, high DNR, FC OFF
S8	Fast WDRC, high DNR, FC ON

Note. WDRC = wide dynamic range compression; DNR = digital noise reduction; FC = frequency compression.

Test Location

Lab

The experiment was originally designed to be completed in a laboratory setting. Participants were seated in a double-walled sound-treated booth. A custom MATLAB program on a Windows computer was used to control the stimulus presentation and record the responses. The program drew from preprocessed stimuli (customized to the participant's audiogram as described under "Hearing Aid Processing") housed on the computer. The presentation of stimuli and responses were controlled by the participant using a graphical user interface (GUI) on the computer screen. Signals were routed to Sennheiser HD 25 headphones via an audio interface (M-Audio M-Trak 8). The maximum output through this system, measured for a 20-s long speech signal consisting of a concatenated string of IEEE sentences, digitally scaled to a peak amplitude of ± 1 was 105.4 dBA. Total harmonic distortion (THD) through this system at 500, 800, and 1600 Hz with amplification for each participant was < 2%. THD at a given frequency was calculated using the formula shown below (American National Standards Institute, 2014). Headphone placement was verified by the experimenter. The lab version of the experiment was completed prior to the COVID-19 pandemic.

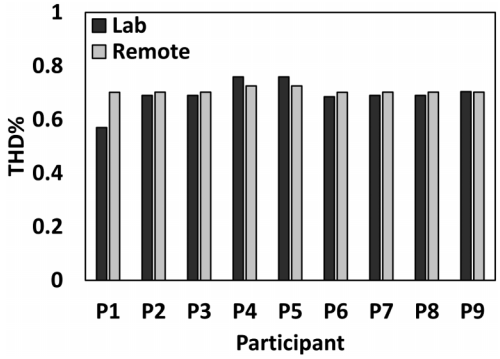
$$THD = 100 * \frac{\sqrt{P_1^2 + P_2^2}}{P_0^2}, \quad (1)$$

where P_0 is the sound pressure of the fundamental, whereas P_1 and P_2 pertain to the sound pressure of the first and second harmonics, respectively.

Remote

Participants completed the remote version of the experiment in their homes during the COVID-19 pandemic. Participants were provided with a Surface Go 2 Tablet (Intel Pentium CPU 4425Y, 1.7 GHz, 4 GB RAM, 64-bit) and the calibrated Sennheiser HD25 headphones. The maximum output through this system, measured for a 20-s long speech signal, consisting of a concatenated string of IEEE sentences, digitally scaled to a peak amplitude of ±1 was 90 dBA. Again, THD through this system at 500, 800, and 1600 Hz with amplification for each participant was < 2%. THD was calculated using (1). Figure 2 shows the three frequency average of THD per participant for the lab and remote systems. For the experiment, the custom MATLAB program from the laboratory version was packaged and deployed as an executable application onto the tablet. Like the lab version, the application drew from preprocessed stimuli housed locally on the tablet. The presentation of stimuli and responses were controlled by the participant using a GUI on the tablet's touchscreen. In addition to the preference data, the timestamp at the end of each trial was also saved. Participants were instructed to seat themselves in a quiet place free from any distractions and were provided with detailed written instructions to operate the tablet. Additional instructions were provided remotely over the phone as needed.

Figure 2. Average total harmonic distortion (THD%) across 500, 800, and 1600 Hz with amplification for individual participants. Dark and light colored bars indicate the THD measured for lab versus remote equipment.



For the remote version of the experiment, two headphone screeners were included as part of the executable. The first headphone screener was designed according to Woods et al. (2017) to ensure that the participant was listening through the headphones. The participant heard three 1-s-long tones generated at 200 Hz each, presented 0.5 s apart. Two out of three tones were presented at 70 dB SPL, and the third tone was presented at 64 dB SPL. Out of the two 70 dB SPL tones, one tone was presented 180° out of phase across the stereo channels. The participant's task was to identify the softest tone out of the three using a three-alternate forced-choice paradigm. If the participant listened through the loudspeakers instead of headphones, the tone that is 180° out of phase would be attenuated and the listener may be unable to identify the softest tone. The participant had to correctly identify five out of six trials to pass the screener. If they failed the screener, they were instructed to verify the headphone placement and repeat the task until they passed.

The second headphone screener was designed to verify that the participant wore the right and left headphones on their right and left ears, respectively (Ellis & Souza, 2020). The screener played a 1-s-long noise burst at 70 dB SPL randomly through the right or left headphone. The participant was instructed to identify which side the noise was heard. The participant passed the screener if they were able to correctly identify six out of six trials. If they failed the screener, they were instructed to verify the headphone placement and repeat the task until they passed.

Equipment was delivered to the participant either by shipping, a socially distanced pickup/drop-off at the participant's residence, or a curbside pickup/drop-off from the Northwestern University campus depending on the participant's convenience. Strict infection control protocols were followed including the use of personal protective equipment by the researcher as well as thorough disinfection of the experimental equipment before and after delivery and pickup. Data were stored on the tablet with a unique code in a hidden folder and uploaded to a secure lab server immediately upon tablet return. Data were then wiped from the tablet before preparing the equipment for the next participant.

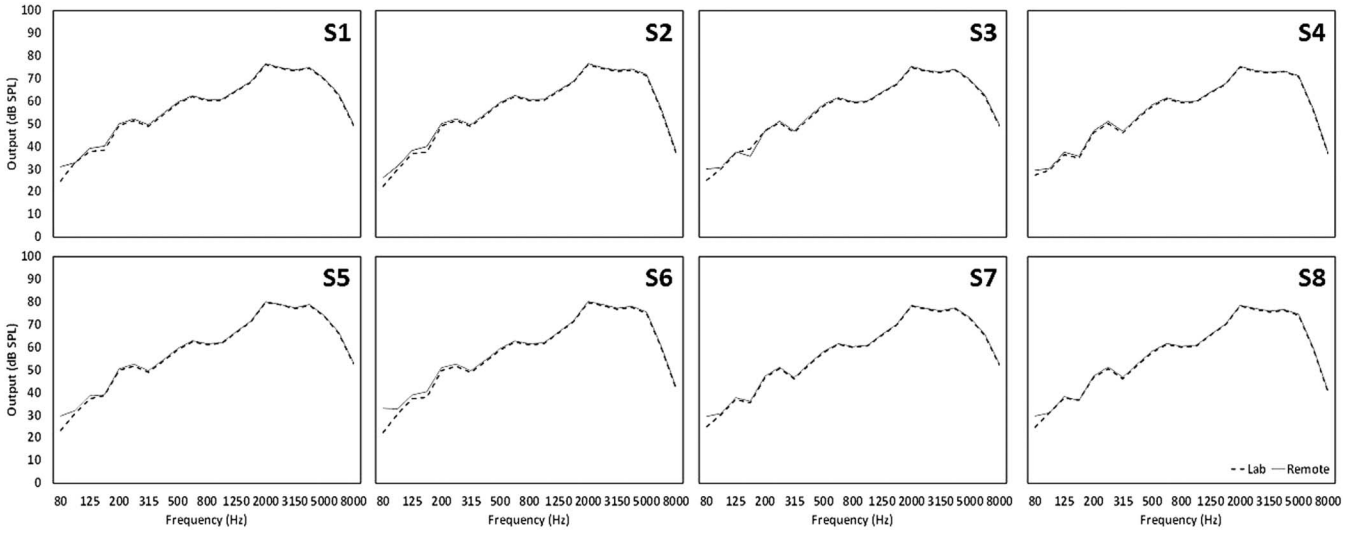
Figure 3 shows the frequency-specific output levels for each hearing aid processing condition, measured using the calibrated lab and remote equipment (right headphone) for a single representative stimulus customized to the average audiogram (see Figure 1). All headphone measurements for Figures 2 and 3 were made with a Bruel &Kjaer 2250 Type I sound-level meter and a 4144 1″ pressure microphone for supra-aural headphones enclosed in a GRAS RA0075 6-cc coupler. Calibrated output levels for both versions were within ± 1 dB between 250 and 8000 Hz. Calibrated output levels for the left headphone were identical to the right headphone and are not shown here.

Procedures

Procedures for the preference experiment were identical for both the lab and remote versions. Participants were instructed to imagine a noisy situation (such as a restaurant) in which they wanted to communicate with a speaker in front of them (i.e., the target sentence), amidst multiple talkers around them. Participants were instructed to use this scenario to make preference judgments for the entire duration of the experiment. Each stimulus (processed target sentence + multitalker babble) represented a "Hearing Aid." Participants heard pairs of stimuli and were instructed to select the "Hearing Aid" that they preferred for listening to the speaker in front of them. Both stimuli in a pair had the same target sentence + babble to minimize confounds due to intelligibility differences. The order of presentation of SNRs was counterbalanced between the lab and remote versions. Sentences were randomized across trials. Each pair of "Hearing Aids" was presented twice, and the order of presentation of stimulus pairs was counterbalanced within a block. Participants provided their responses on a GUI using either a mouse click (lab version) or the touchscreen (remote version). Each round of preference judgments lasted for ~120 min.

Candidacy assessment was completed at the beginning of the lab version. All participants were reconsented for the remote version of the experiment using secure REDCap electronic data capture tools hosted at Northwestern University. Consent forms were e-mailed to the participant and could only be accessed using a unique code provided by the experimenter. Participants completed the remote version of the experiment within a span of 12 months (range: 7–11 months) from the lab version. The exception was one participant who completed the remote version after 15 months as she preferred to wait until after vaccination. There was no significant change in the participant's hearing status between the two sessions. No

Figure 3. Output levels measured from the right headphone across 1/3rd octave frequencies for a sample stimulus at +3 dB SNR. Solid and dashed lines represent the levels measured from the remote and lab equipment, respectively. Frequency shaping is based on the average audiogram shown in Figure 1. Each panel is a separate hearing aid processing condition (S1–S8; see Table 1).



assumption was made regarding the speech intelligibility or quality of the signal for the listener as this study was focused on eliciting preference.

Statistical Analyses

Statistical analyses were conducted using McNemar's test of preference (Fagerland et al., 2013) to determine the association between "Hearing Aid" preference and experiment versions (lab vs. remote). This test is comparable to a paired t test and is appropriate for determining differences between two dependent groups for a categorical outcome such as preference. A stepdown Bonferroni adjustment was applied for multiple comparisons (Westfall et al., 2010). Additional analyses were conducted to determine the agreement between the two experiment versions by computing the percentage of agreement in preference and kappa scores for each "Hearing Aid" (see Table 2). The kappa score is considered a measure of "true" agreement as it accounts for agreement beyond chance alone (Sim & Wright, 2005). For the kappa score, the following standards of strength apply: < 0 = no agreement, .01–.20 = slight, .21–.40 = fair, .41–.60 = moderate, .61–.80 = substantial, and .81–1 = almost perfect agreement (Cohen, 1960; McHugh, 2012).

Results

Figure 4 shows the observed proportion of preference for both versions of the experiment across each of the eight "Hearing Aids." Note that the y-axis shows the proportion of preference for a given "Hearing Aid" compared to all other "Hearing Aids" at a given SNR. McNemar's test of preference found no statistically significant association between "Hearing Aid" preference and experiment version (i.e., lab vs. remote) across the majority of the "Hearing Aid" settings at both SNRs ($p > .05$). The only exception was "Hearing Aid" S5 or the combination of Fast WDRC, High DNR, and FC OFF, which resulted in a higher proportion of preference with the remote version compared to the lab version at 8 dB SNR ($\chi^2 = 12.812$, $p < 0.01$). The reason for this exception is not clear and requires further investigation. Results of the McNemar test are shown in Table 2.

There was a moderate percentage agreement (60%–70%), and the kappa score indicated fair agreement (0.21 to 0.40) between the two versions for at least five "Hearing Aids" at 3 dB SNR and seven "Hearing Aids" at 8 dB SNR.

Discussion

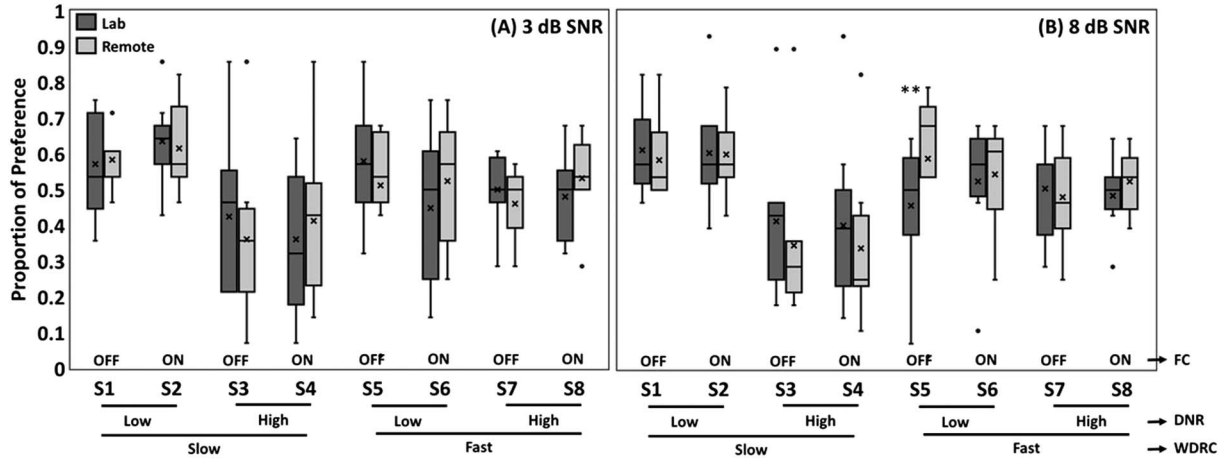
The goal of this study was to determine the feasibility of a remote data collection method to elicit preference for hearing aid signal processing using an auditory paired-comparisons task. The remote version was conducted in a quiet room in the participant's home, whereas the lab version was conducted in a double-walled sound-treated booth. The remote version used a tablet-based platform for conducting the experiment along with calibrated headphones for stimulus delivery. The lab version used a desktop computer for conducting the experiment, and calibrated headphones were connected to an audio interface for stimulus delivery. The outcome measure (preference) was compared between the lab and remote versions for a group of nine listeners.

Table 2. Results of McNemar's test (χ^2), percentage agreement, and kappa scores comparing the remote and lab versions of the experiment.

SNR	"Hearing Aid"	χ^2	Adjusted p	% Agreement	Kappa, 95% CI [LL, UL]
3 dB	S1	0.101	1.000	62.14%	0.276 [0.157, 0.396]
	S2	0.263	1.000	62.14%	0.196 [0.073, 0.319]
	S3	3.460	1.000	69.29%	0.387 [0.272, 0.502]
	S4	1.707	.881	61.07%	0.174 [0.052, 0.297]
	S5	2.919	1.000	59.29%	0.211 [0.092, 0.331]
	S6	3.574	.881	58.57%	0.202 [0.083, 0.322]
	S7	1.220	1.000	66.07%	0.349 [0.234, 0.465]
	S8	1.817	1.000	62.86%	0.264 [0.145, 0.382]
8 dB	S1	0.551	1.000	63.21%	0.267 [0.146, 0.387]
	S2	0.012	1.000	66.07%	0.313 [0.194, 0.432]
	S3	3.176	.971	62.14%	0.237 [0.116, 0.358]
	S4	2.844	1.000	63.21%	0.236 [0.115, 0.358]
	S5	12.812	**.005**	65.71%	0.336 [0.224, 0.447]
	S6	0.269	1.000	62.86%	0.259 [0.140, 0.378]
	S7	0.367	1.000	60.36%	0.223 [0.102, 0.343]
	S8	0.980	1.000	58.93%	0.192 [0.071, 0.312]

Note. p values are reported for the McNemar's test, and values that are significant are highlighted in bold. SNR = signal-to-noise ratio; CI = confidence interval; LL = lower limit; UL = upper limit.

Figure 4. Boxplots comparing proportion of preference (*y*-axis) between in-lab (dark bars) versus remote (light bars) data collection methods across "Hearing Aids" (*x*-axis, S1–S8). Each panel shows a different signal-to-noise-ratio (SNR; A = 3 dB, B = 8 dB). Labels on the *x*-axis show the hearing aid signal processing attributes and levels constituting each "Hearing Aid." Asterisks (**) represent $p < .01$. FC = frequency compression; DNR = digital noise reduction; WDRC = wide dynamic range compression.



A McNemar's test showed no significant association between experiment version and preference, suggesting that listeners' preferences for hearing aid settings were not influenced by the data collection method. Note that while we did not detect many statistically significant differences in preference between locations, the relatively small number of participants means that such analyses are underpowered. Indeed, for pairwise preferences between two "Hearing Aids," we had 80% power to detect very large differences in preference probability (i.e., differences > 70%). However, percentage agreement and kappa statistics (a measure of "true" agreement, beyond chance alone) indicated moderate/fair agreement for most "Hearing Aids" with reasonably wide confidence intervals (CIs; 95% CI widths roughly 0.2, 0.25). Thus, the data suggest agreement that is likely far from "strong" or perfect, but also consistently better than zero. The distribution of preference in Figure 4 suggests no systematic differences between the two experiment versions for most of the hearing aid settings. Moreover, the output levels across conditions did not differ between the two versions (see Figure 3). Therefore, it is likely that the relatively low agreement between the versions was influenced by within-subject variability in the outcome measure (preference) rather than the data collection methods themselves. There is evidence that while a paired comparison task for determining hearing aid–related preference is reliable, factors such as age and hearing loss may introduce some within-subject variability (see the work of Amlani and Schafer, 2009, for a review). Due to the small sample size, this study does not have adequate power to characterize individual variability in preference. This variability will need to be accounted for through statistical methods in the main study with a larger group of participants.

One of the concerns with any remote experimentation is insufficient headroom to provide adequate amplification for individuals with hearing loss. Although the equipment used in the remote version of this study had lower headroom than the lab version, it was sufficient to provide the necessary amplification for the participants in the mild to moderately severe hearing loss range. Of course, for studies conducted on individuals with more severe hearing losses, a different set of headphones or an in-line amplifier may need to be considered to increase the headroom.

Participants were able to successfully use the tablet-based platform to carry out this experiment. While we did not formally evaluate the performance differences between the lab-based and remote equipment, anecdotal reports from participants revealed a few technical issues. There was one instance of a loss of headphone detection at the beginning of the experiment. This was resolved by reinstructing the participant over the phone to plug in the headphones securely. Use of the headphone screeners at the beginning of each experiment (repeated if the participant returned to the experiment after a break) ensured that the headphones were always appropriately connected for the task. Most of the participants reported that the executable application was slow to launch (i.e., ~30–45 s slower than the lab computer). This could be from the combination of the MATLAB runtime environment and the low processing power of the tablet used in the remote version of the experiment. A tablet with low processing power was chosen to conserve costs and factor in replacements in the event of loss and damage during remote delivery.

Time elapsed within (time taken by the MATLAB program to retrieve and play a pair of preprocessed stimuli after pressing "Play") and between trials (time

taken by the MATLAB program to save the response and switch to the next trial) was computed for a single block of 64 trials for a dummy stimulus set. With the lab-based equipment, average time elapsed within and between trials was 12.22 s (SD = 0.60) and 0.20 s (SD = 0.40), respectively. With the remote equipment, average time lapsed within and between trials was 12.34 s (SD = 0.56) and 0.39 s (SD = 0.06) respectively. Thus, once the GUI was launched, there was negligible difference in the experiment speed within and between trials between the two versions. All participants completed each of the versions of the experiment within 2 hr, including breaks.

While we did not measure noise levels in the participants, homes where the remote experiment was completed, recent studies have shown that ambient noise levels in home environments are fairly low (Kopun et al., 2021; Ramos et al., 2022) and not likely to interfere with testing using supra-aural headphones. Moreover, these listeners did not have normal hearing and the noisy speech was amplified, such that audibility would have been determined by the stimulus SNR. That is, the experiment was not operating at threshold/low levels to be impacted by the background noise in the environment.

Finally, this study established the feasibility of a remote testing method for audiological experiments using a custom MATLAB executable application on a tablet and calibrated headphones. For studies involving customized hearing aid signal processing for listeners with hearing loss, this is a reasonable alternative to conducting studies in the lab, which allows for maintaining experimental control while providing a socially distanced and flexible test setup. While the primary purpose of remote testing was to overcome the barriers posed by COVID-19 pandemic, these methods can be applied to other situations where commute or accessibility to the lab may be limited.

Acknowledgments

This work was supported by the American Speech and Hearing Foundation New Investigators Grant (awarded to V.R.). Portions of this work were presented at the American Speech-Language-Hearing Association 2021 Convention. The authors thank Jacob Schauer for assistance with statistical analysis, and Kendra Marks and Magda Wisniewska for assistance with data collection.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author (V.R.), upon reasonable request.

References

Acoustical Society of America Psychological and Physiological Acoustics Task Force on Remote Testing. (2020). *Remote testing Wiki: ASA P&P task force on remote testing*. https://www.spatialhearing.org/remotetesting/Main/HomePage

Alexander, J. M. (2016a). Hearing aid delay and current drain in modern digital devices. *Canadian Audiologist, 3*(4). https://canadianaudiologist.ca/hearing-aid-delay-feature/

Alexander, J. M. (2016b). Nonlinear frequency compression: Influence of start frequency and input bandwidth on consonant and vowel recognition. *The Journal of the Acoustical Society of America, 139*(2), 938–957. https://doi.org/10.1121/1.4941916

Alexander, J. M., & Masterson, K. (2015). Effects of WDRC release time and number of channels on output SNR and speech recognition. *Ear and Hearing, 36*(2), e35–e49. https://doi.org/10.1097/AUD.0000000000000115

American National Standards Institute. (2014). *Specification of hearing aid characteristics (ANSI/ASA S3.22)*.

Amlani, A. M., & Schafer, E. C. (2009). Application of paired-comparison methods to hearing AIDS. *Trends in Amplification, 13*(4), 241–259. https://doi.org/10.1177/1084713809352908

Arehart, K. H., Chon, S. H., Lundberg, E. M. H., Harvey, L. O., Jr., Kates, J. M., Anderson, M. C., Rallapalli, V. H., & Souza, P. E. (2022). A comparison of speech intelligibility and subjective quality with hearing-aid processing in older adults with hearing loss. *International Journal of Audiology, 61*(1), 46–58. https://doi.org/10.1080/14992027.2021.1900609

Arehart, K. H., Souza, P., Kates, J. M., Lunner, T., & Pedersen, M. S. (2015). Relationship among signal fidelity, hearing loss, and working memory for digital noise suppression. *Ear and Hearing, 36*(5), 505–516. https://doi.org/10.1097/AUD.0000000000000173

Bentler, R., & Chiou, L. K. (2006). Digital noise reduction: An overview. *Trends in Amplification, 10*(2), 67–82. https://doi.org/10.1177/1084713806289514

Berouti, M., Schwartz, R., & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise [Paper presentation]. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 4, pp. 208–211), Washington, DC. IEEE.

Bisgaard, N., Vlaming, M. S., & Dahlquist, M. (2010). Standard audiograms for the IEC 60118-15 measurement procedure. *Trends in Amplification, 14*(2), 113–120. https://doi.org/10.1177/1084713810379609

Bridges, J. F. P., Lataille, A. T., Buttorff, C., White, S., & Niparko, J. K. (2012). Consumer preferences for hearing aid attributes: A comparison of rating and conjoint analysis methods. *Trends in Amplification, 16*(1), 40–48. https://doi.org/10.1177/1084713811434617

Brons, I., Houben, R., & Dreschler, W. A. (2015). Acoustical and perceptual comparison of noise reduction and compression in hearing aids. *Journal of Speech, Language, and Hearing Research, 58*(4), 1363–1376. https://doi.org/10.1044/2015_JSLHR-H-14-0347

Cohen, I., & Berdugo, B. (2002). Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Processing Letters, 9*(1), 12–15. https://doi.org/10.1109/97.988717

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Dillon, H. (2012). *Hearing aids* (2nd ed.). Thieme.

Ellis, G. E., & Souza, P. (2020). *Effects of hearing aid processing on speech intelligibility in virtual restaurant settings*. 177th Meeting of the Acoustical Society of America, Chicago, Illinois. https://doi.org/10.1121/1.5147755

Eskenazi, M., Levow, G.-A., Meng, H., Parent, G., & Suendermann, D. (2013). *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. Wiley. https://doi.org/10.1002/9781118541241

Fagerland, M. W., Lydersen, S., & Laake, P. (2013). The McNemar test for binary matched-pairs data: Mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology, 13*(1), 91. https://doi.org/10.1186/1471-2288-13-91

Fortune, T. W., & Preves, D. A. (1992). Hearing aid saturation and aided loudness discomfort. *Journal of Speech and Hearing Research, 35*(1), 175–185. https://doi.org/10.1044/jshr.3501.175

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., & Duda, S. N. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics, 95,* 103208. https://doi.org/10.1016/j.jbi.2019.103208

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Jenstad, L. M., & Souza, P. (2005). Quantifying the effect of compression hearing aid release time on speech acoustics and intelligibility. *Journal of Speech, Language, and Hearing Research, 48*(3), 651–667. https://doi.org/10.1044/1092-4388(2005/045)

Kates, J. M. (2017). Modeling the effects of single-microphone noise-suppression. *Speech Communication, 90,* 15–25. https://doi.org/10.1016/j.specom.2017.04.004

Kates, J. M., Arehart, K. H., & Harvey, L. O., Jr. (2019). Integrating a remote microphone with hearing-aid processing. *The Journal of the Acoustical Society of America, 145*(6), 3551–3566. https://doi.org/10.1121/1.5111339

Keidser, G., Dillon, H., Carter, L., & O'Brien, A. (2012). NAL-NL2 empirical adjustments. *Trends in Amplification, 16*(4), 211–223. https://doi.org/10.1177/1084713812468511

Kim, G., & Loizou, P. C. (2011). Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms. *The Journal of the Acoustical Society of America, 130*(3), 1581–1596. https://doi.org/10.1121/1.3619790

Kochkin, S. (2007). MarkeTrak VII. *Hearing Journal, 60*(4), 24–51. https://doi.org/10.1097/01.HJ.0000285745.08599.7f

Kopun, J. G., Turner, M., Harris, S. E., Kamerer, A. M., Neely, S. T., & Rasetshwane, D. M. (2021). Evaluation of Remote Categorical Loudness Scaling. *American Journal of Audiology,* 1–12.

Kuk, F. K. (2002). Paired comparisons as a fine-tuning tool in hearing aid fittings. In M. Valente (Ed.), *Strategies for selecting and verifying hearing aid fittings* (2nd ed., pp. 125–150). Thieme.

Margolis, R. H., Killion, M. C., Bratt, G. W., & Saly, G. L. (2016). Validation of the Home Hearing Test™. *Journal of the American Academy of Audiology, 27*(5), 416–420. https://doi.org/10.3766/jaaa.15102

McCormack, A., & Fortnum, H. (2013). Why do people fitted with hearing aids not wear them? *International Journal of Audiology, 52*(5), 360–368. https://doi.org/10.3109/14992027.2013.769066

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

McPherson, M. J., & McDermott, J. H. (2020). Time-dependent discrimination advantages for harmonic sounds suggest efficient coding for memory. *Proceedings of the National Academy of Sciences, 117*(50), 32169–32180. https://doi.org/10.1073/pnas.2008956117

Moore, B. C. (2008). The choice of compression speed in hearing aids: Theoretical and practical considerations and the role of individual differences. *Trends in Amplification, 12*(2), 103–112. https://doi.org/10.1177/1084713808317819

Mosley, C. L., Langley, L. M., Davis, A., McMahon, C. M., & Tremblay, K. L. (2019). Reliability of the Home Hearing Test: Implications for public health. *Journal of the American Academy of Audiology, 30*(3), 208–216. https://doi.org/10.3766/jaaa.17092

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society, 53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Panfili, L. M., Haywood, J., McCloy, D. R., Souza, P. E., & Wright, R. A. (2017). *University of Washington/Northwestern University (UW/NU) Corpus 2.0*. https://depts.washington.edu/phonlab/projects/uwnu.php

Rallapalli, V. H., Mueller, A., Appleton, R., & Souza, P. E. (2018). Survey of hearing aid signal processing features across manufacturers. *Journal of the American Academy of Audiology*, 29(2), 118–124. https://doi.org/10.3766/jaaa.16107

Ramos, J. R. V., Marks, K., & Souza, P. (2022). *Ambient noise measurements using smartphone SLM applications in home environments* [Poster presentation]. 62nd Illinois Speech and Hearing Association Annual Convention, Rosemont, IL.

Ricketts, T., Bentler, R., & Mueller, G. (2019). *Essentials of modern hearing aids: Selection, fitting, and verification*. Plural.

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., & Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics, 17*(3), 225–246. https://doi.org/10.1109/TAU.1969.1162058

Shapiro, M. L., Norris, J. A., Wilbur, J. C., Brungart, D. S., & Clavier, O. H. (2020). TabSINT: Open-source mobile software for distributed studies of hearing. *International Journal of Audiology, 59*(Suppl. 1), S12–S19. https://doi.org/10.1080/14992027.2019.1698776

Schoeffler, M., Bartoschek, S., Stöter, F. R., Roess, M., Westphal, S., Edler, B., & Herre, J. (2018). webMUSHRA—A comprehensive framework for web-based listening tests. *Journal of Open Research Software, 6*(1). https://doi.org/10.5334/jors.187

Shen, J., Anderson, M., Arehart, K. H., & Souza, P. (2016). Using cognitive screening tests in audiology. *American Journal of Audiology, 25*(4), 319–331. https://doi.org/10.1044/2016_AJA-16-0032

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Smeds, K., Wolters, F., & Rung, M. (2015). Estimation of signal-to-noise ratios in realistic sound scenarios. *Journal of the American Academy of Audiology, 26*(2), 183–196. https://doi.org/10.3766/jaaa.26.2.7

Souza, P. (2002). Effects of compression on speech acoustics, intelligibility, and sound quality. *Trends in Amplification, 6*(4), 131–165. https://doi.org/10.1177/108471380200600402

Souza, P., Arehart, K., Schoof, T., Anderson, M., Strori, D., & Balmert, L. (2019). Understanding variability in individual response to hearing aid signal processing in wearable hearing aids. *Ear and Hearing, 40*(6), 1280–1292. https://doi.org/10.1097/AUD.0000000000000717

Souza, P., Arehart, K. H., Kates, J. M., Croghan, N. B., & Gehani, N. (2013). Exploring the limits of frequency lowering. *Journal of Speech, Language, and Hearing Research, 56*(5), 1349–1363. https://doi.org/10.1044/1092-4388(2013/12-0151)

Souza, P., Arehart, K. H., & Neher, T. (2015). Working memory and hearing aid processing: Literature findings, future directions, and clinical applications. *Frontiers in Psychology, 6*(6), 1894. https://doi.org/10.3389/fpsyg.2015.01894

Souza, P., Arehart, K. H., Shen, J., Anderson, M., & Kates, J. M. (2015). Working memory and intelligibility of hearing-aid processed speech. *Frontiers in Psychology, 6,* 526. https://doi.org/10.3389/fpsyg.2015.00526

Strori, D., Bradlow, A. R., & Souza, P. E. (2020). Recognition of foreign-accented speech in noise: The interplay between talker intelligibility and linguistic structure. *The Journal of the Acoustical Society of America, 147*(6), 3765–3782. https://doi.org/10.1121/10.0001194

Virag, N. (1999). Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on Speech and Audio Processing, 7*(2), 126–137.

Walker, A., & Campbell-Kibler, K. (2015). Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. *Frontiers in Psychology, 6,* 546. https://doi.org/10.3389/fpsyg.2015.00546

Waz, S., & Chubb, C. (2020). How do listeners use context frequencies in tone-scramble tasks? Evidence from a web-based experiment. *The Journal of the Acoustical Society of America, 148*(4), 2715–2716. https://doi.org/10.1121/1.5147529

Westfall, P. H., Troendle, J. F., & Pennello, G. (2010). Multiple McNemar tests. *Biometrics, 66*(4), 1185–1191. https://doi.org/10.1111/j.1541-0420.2010.01408.x

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Wu, Y. H., Stangl, E., Chipara, O., Hasan, S. S., Welhaven, A., & Oleson, J. (2018). Characteristics of real-world signal to noise ratios and speech listening situations of older adults with mild to moderate hearing loss. *Ear and Hearing, 39*(2), 293–304. https://doi.org/10.1097/AUD.0000000000000486

Yu, A. C., & Lee, H. (2014). The stability of perceptual compensation for coarticulation within and across individuals: A cross-validation study. *The Journal of the Acoustical Society of America, 136*(1), 382–388. https://doi.org/10.1121/1.4883380

Zahorik, P. (2009). Perceptually relevant parameters for virtual listening simulation of small room acoustics. *The Journal of the Acoustical Society of America, 126*(2), 776–791. https://doi.org/10.1121/1.3167842

Appendix

Hearing Aid Simulation Settings

Feature	Description/setting
Version	4
Filterbank	Linear-phase FIR filters; group delay = ½ filter length & independent of frequency; filter length = 16 ms; filter output adjusted for filter onset and offset transients to result in net zero-time delay
Sampling rate	44100 Hz
Microphone response	Behind-the-ear hearing aid shell; resonance peak at 5 kHz; designed using an IIR filter
Receiver response	Resonance peak at 2 kHz; designed using an IIR filter
Processing delay	10 ms
Vent	Fully occluded
Frequency compression	5-pole Butterworth high-pass and low-pass filters used to create a 2-band system divided by a cut-off (start) frequency and transformed into IIR filters; no processing applied to low frequency band; sinusoidal modeling applied to high frequency band using the sequence described in Souza et al. (2013)
Digital noise reduction	16-ms raised cosine window with 50% overlap used to segment filterbank outputs; Noise estimation is as described in "Method" with a 20-dB threshold (above valley) for classifying signal as speech; initial estimate of noise power is obtained for the first 50 ms of the noisy speech input signal; Wiener filtering is applied according to the formula and parameters in Berouti et al. (1979) and Virag (1999)
Wide dynamic range compression	Lower and upper compression thresholds are applied to all 12 channels; NAL-NL2 parameters include listener gender, compression time constants, and binaural fit; gains are chosen for the nearest MMSE match of the participant's audiogram to the ISMADHA standard audiograms (Bisgaard et al., 2010)

Note. FIR = finite impulse response; IIR = infinite impulse response; MMSE = minimum mean squared error; ISMADHA = International Standards for Measuring Advanced Digital Hearing Aids.